Projeto 4: Prevendo o Risco de Calote

Entendimento de negócios e dados

A gerência do banco necessita decidir, no prazo máximo de uma semana, quais clientes, dentre uma lista de 500 solicitações, devem ter seu pedido de empréstimo aprovado. Para tanto, será necessário prever se o cliente irá quitar o empréstimo conforme as condições contratadas ou não. Por fim, deverá ser informado quantos clientes devem ter seu pedido de crédito aprovado.

Os dados necessários à construção de um modelo preditivo abrangem resultados de empréstimos anteriores, valor do empréstimo concedido, duração do empréstimo e renda familiar, essencialmente. Outras variáveis, como idade e estabilidade no emprego atual também são desejáveis.

No caso em análise, estão disponíveis os seguintes dados, relativos a clientes antigos do banco e também aos potenciais novos clientes: “Credit-Application-Result”(que consiste na variável-alvo e, portanto, se referem aos clientes antigos), “Account-Balance”, “Duration-of-Credit-Month”, “Payment-Status-of-Previous-Credit”, “Purpose”, “Credit-Amount”, “Value-Savings-Stocks”, “Length-of-current-employment”, “Instalment-per-cent”, “Guarantors”, “Duration-in-Current-address”, “Most-valuable-available-asset”, “Age-years”, “Concurrent-Credits”, “Type-of-apartment”, “No-of-Credits-at-this-Bank”, “Occupation”, “No-of-dependents”, “Telephone”, “Foreign-Worker”.

Como o problema que se apresenta envolve a predição de um resultado e muitas informações sobre antigos clientes estão disponíveis, o modelo preditivo buscado é considerado como rico em dados. Por outro lado, o resultado buscado – crédito aprovado ou não aprovado – envolve a classificação entre duas possíveis categorias e sua predição não está baseada em tempo. Assim sendo, o modelo preditivo a ser utilizado é de classificação binário.

Construção do Conjunto de Treinamento

“Telephone” não possui relação com a variável-alvo; logo não há razão lógica para incluí-la no modelo.

A matriz de correlação a seguir demonstra a inexistência de correlação interna forte – acima de 0,70 – entre as variáveis numéricas, o u seja, todas podem ser usadas como preditivas.

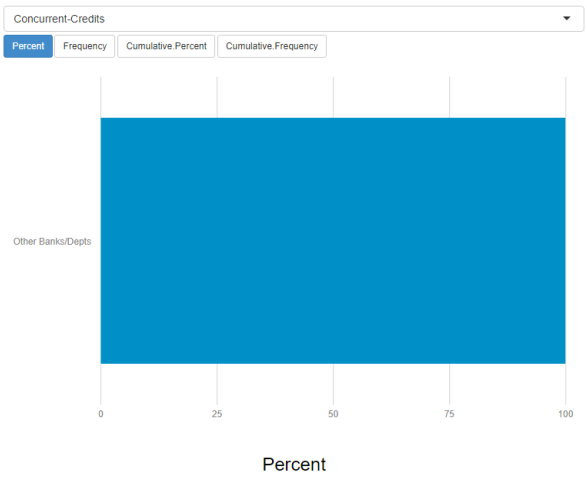
Full Correlation Matrix

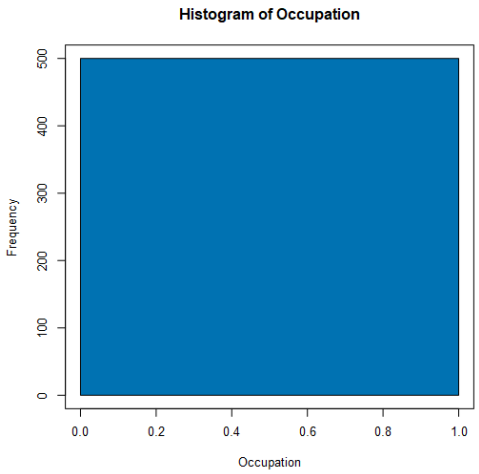
	Credit.Application.Result.num	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years
Credit.Application.Result.num	1.0000000	-0.2043168	-0.2009899	-0.0653449	-0.1379166	0.0567366
Duration.of.Credit.Month	-0.2043168	1.0000000	0.5704408	0.0795146	0.3047342	-0.0663189
Credit.Amount	-0.2009899	0.5704408	1.0000000	-0.2856309	0.3277621	0.0686430
Instalment.per.cent	-0.0653449	0.0795146	-0.2856309	1.0000000	0.0781104	0.0405397
Most.valuable.available.asset	-0.1379166	0.3047342	0.3277621	0.0781104	1.0000000	0.0854367
Age.years	0.0567366	-0.0663189	0.0686430	0.0405397	0.0854367	1.0000000
Type.of.apartment	-0.0218604	0.1531405	0.1686831	0.0829360	0.3796504	0.3330748
No.of.dependents	-0.0387889	-0.0604413	0.0055003	-0.1164661	0.0507817	0.1177351
Foreign.Worker	0.0056897	-0.1064163	0.0318954	-0.1182555	-0.1405878	-0.0032847

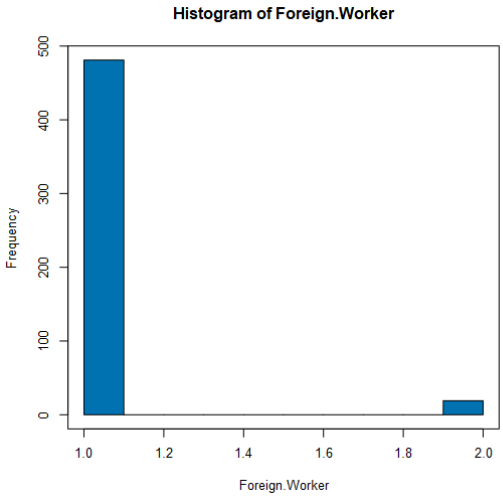
	Type.of.apartment	No.of.dependents	Foreign.Worker
Credit.Application.Result.num	-0.0218604	-0.0387889	0.0056897
Duration.of.Credit.Month	0.1531405	-0.0604413	-0.1064163
Credit.Amount	0.1686831	0.0055003	0.0318954
Instalment.per.cent	0.0829360	-0.1164661	-0.1182555
Most.valuable.available.asset	0.3796504	0.0507817	-0.1405878
Age.years	0.3330748	0.1177351	-0.0032847
Type.of.apartment	1.0000000	0.1707221	-0.0968173
No.of.dependents	0.1707221	1.0000000	0.0412103
Foreign.Worker	-0.0968173	0.0412103	1.0000000

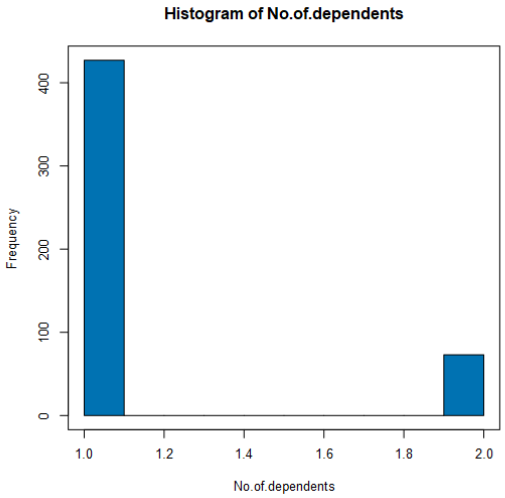
Abaixo os histogramas das variáveis:

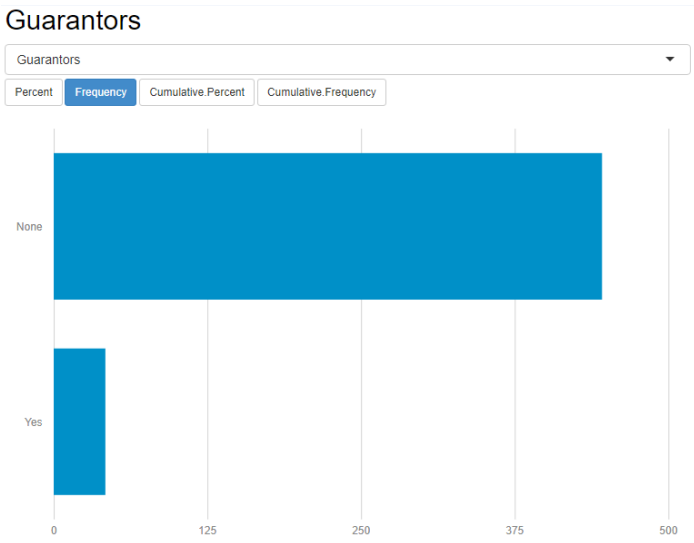
Variáveis removidas por baixa variabilidade: “Concurrent-Credits” e “Occupation” (dados inteiramente uniformes), e “Foreign-Worker”, “Guarantors” e “No-of-dependents” (fortemente enviesadas para um tipo de dado).


Concurrent-Credits
Percent
Frequency
Cumulative.Percent
Cumulative.Frequency
Other Banks/Depts
0
25
50
75
100
Percent

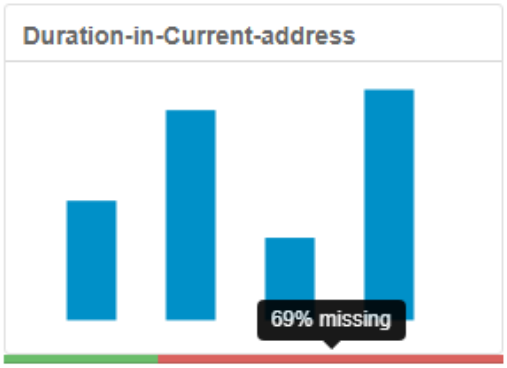

Histogram of Occupation
Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
Occupation


Histogram of Foreign.Worker
Frequency
0
100
200
300
400
500
1.0
1.2
1.4
1.6
1.8
2.0
Foreign.Worker

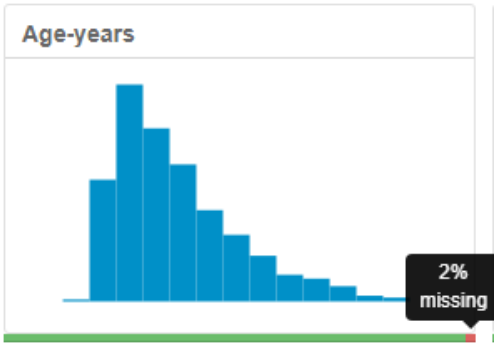

Histogram of No.of.dependents
Frequency
0
100
200
300
400
1.0
1.2
1.4
1.6
1.8
2.0
No.of.dependents


Guarantors
Guarantors
Percent
Frequency
Cumulative.Percent
Cumulative.Frequency
None
Yes
0
125
250
375
500

“Duration-in-Current-address” foi removida, pois apresentou percentual elevado de dados ausentes:



Os dados faltantes de “Age-years” (apenas 2%) foram imputados pela mediana, devido à distribuição assimétrica à direita:



Treinamento dos Modelos de Classificação

Para a investigação de qual a melhor solução para a situação em tela, foram criadas amostras de estimação (70% do conjunto de dados) e de validação (30% do conjunto de dados), tendo sido aplicadas aos seguintes modelos: *regressão logística, árvore de decisão (decision trees), modelo de floresta (forest model) e boosted model.*

No modelo de Regressão Logística, as variáveis “Account-Balance”, “Payment-Status-of-Previous-Credit”, “Purpose”, “Credit-Amount”, “Length-of-current-employment” e “Instalment-per-cent” foram consideradas estatisticamente significativas (presença de asteriscos no demonstrativo do Alteryx).

Report for Logistic Regression Model SP_Credit_log

Basic Summary

Call:
glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial("logit"), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183	*
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566	**
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042	
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618	.
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296	**
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545	
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596	*
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549	*
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Já no modelo de Árvore de Decisão, as variáveis consideradas importantes foram "Account-Balance", "Duration-of-Credit-Month" e "Value-Savings-Stocks". A precisão (accuracy) do modelo com base no conjunto de treinamento foi 78%.

Summary Report for Decision Tree Model DT_Credit

Call:
rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + No.of.Credits.at.this.Bank, data = the.data, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 1e-05, usesurrogate = 0, surrogatestyle = 0)

Model Summary

Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

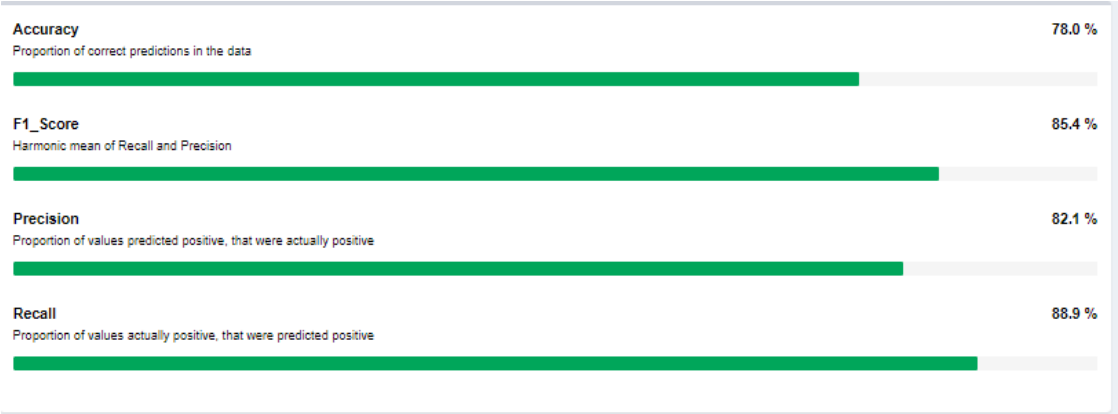
Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

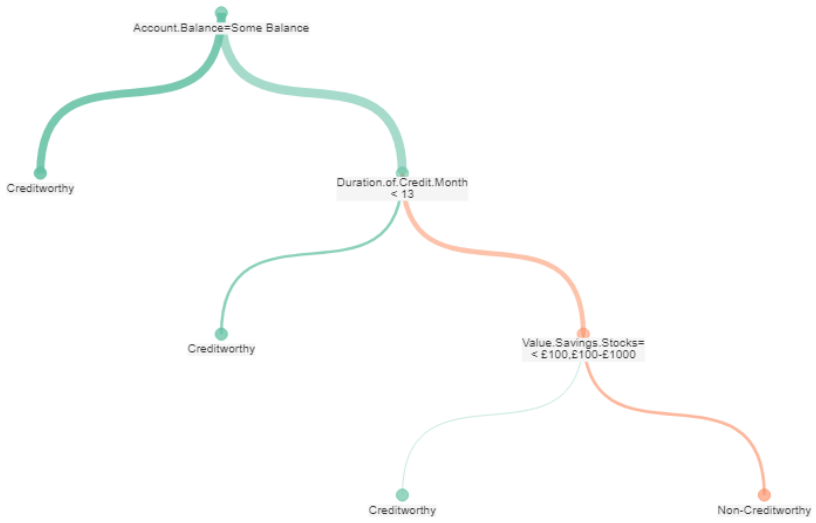
Leaf Summary

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 350 97 Creditworthy (0.7228571 0.2771429)
2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789) *



Actual	Actual Positive	Actual Negative
Predicted Positive	48 (49.5%)	49 (50.5%)
Predicted Negative	28 (11.1%)	225 (88.9%)



Com relação ao Modelo de Floresta, as três variáveis consideradas mais importantes foram “Credit-Amount”, “Age-years” e “Duration-of-Credit-Month”. A taxa de erro de estimação foi 24,6% e a matriz de confusão com base no conjunto de estimativa apresentou erro de 8,7% para “Creditworthy” e 66% para “Non-Creditworthy”.

Basic Summary

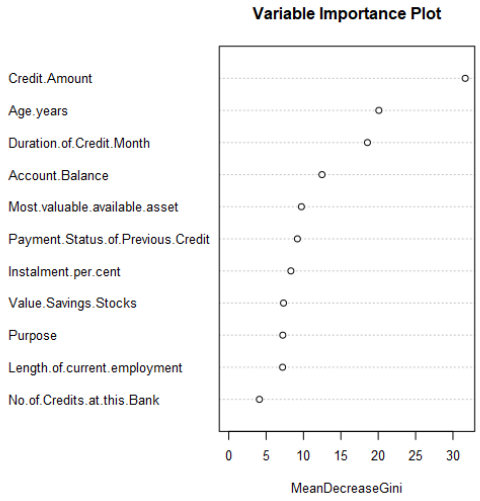
Call:
randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + No.of.Credits.at.this.Bank, data = the.data, ntree = 500, replace = TRUE)

Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3

OOB estimate of the error rate: 24.6%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.66	64	33

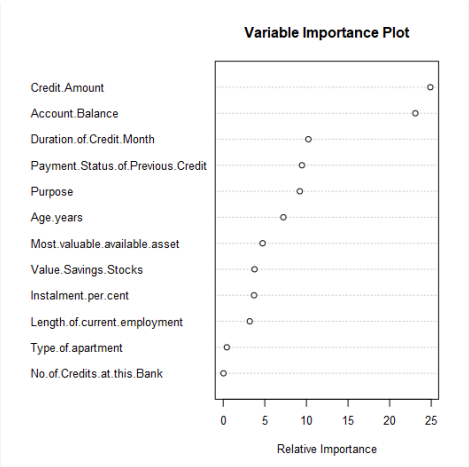


No que tange ao *Boosted Model*, as variáveis consideradas mais significativas foram "Credit-Amount", "Account-Balance", e "Duration-of-Credit-Month".

Report for Boosted Model BM_Credit

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 3940



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.

Para a validação dos modelos construídos, foi utilizada a ferramenta “Model Comparison” do Alteryx juntamente com o conjunto de dados independentes, tendo sido obtidas as seguintes informações:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
SP_Credit_log	0.7600	0.8364	0.7306	0.8762	0.4889
DT_Credit	0.7467	0.8273	0.7054	0.8667	0.4667
FM_Credit	0.8000	0.8707	0.7421	0.9619	0.4222
BM_Credit	0.7933	0.8670	0.7509	0.9619	0.4000

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of DT_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of FM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of SP_Credit_log

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

O *Boosted Model* (BM_Credit) apresentou precisão geral de 79,33%, a segunda melhor. A matriz de confusão correspondente demonstrou uma ocorrência de falso positivo de 27 e falso negativo de 4. Observa-se, em consequência, um viés na predição de “Non-Creditworthy”, que é proporcionalmente mais difícil de se acertar.

O modelo de Árvore de Decisão (DT_Credit) apresentou precisão geral de 74,67%, a mais baixa. De se registrar que uma baixa confiabilidade costuma ocorrer em árvores de decisão devido à sua tendência de superajustar o modelo ao conjunto de dados. A matriz de confusão correspondente demonstrou uma ocorrência de falso positivo de 24 e de falso negativo de 14. Embora a desproporção seja um pouco menor, a predição de “Non-Creditworthy” continua mais difícil.

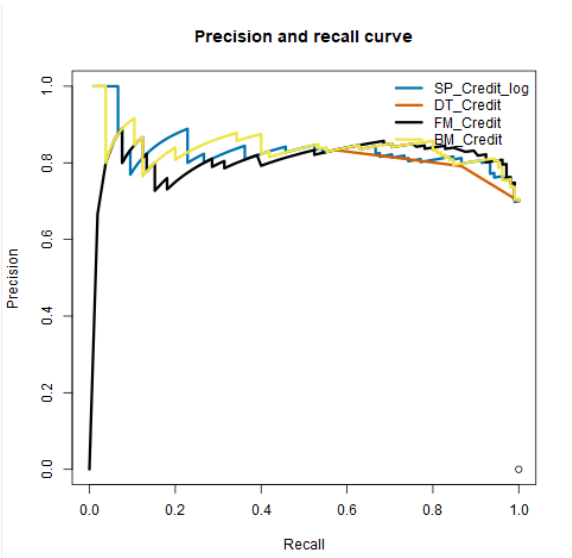
O Modelo de Floresta (FM_Credit) apresentou precisão geral de 80%, a maior entre as apuradas. A matriz de confusão correspondente demonstrou ocorrência de falso positivo de 26 e de falso negativo de 4. O viés na predição de “Non-Creditworthy”, proporcionalmente mais difícil, é um pouco menor que o verificado no *boosted model*.

O modelo de regressão logística (SP_Credit-log) apresentou precisão geral de 76%. A matriz de confusão correspondente demonstrou falso positivo de 23 e falso negativo de 13, o que evidencia um menor viés na predição das duas categorias.

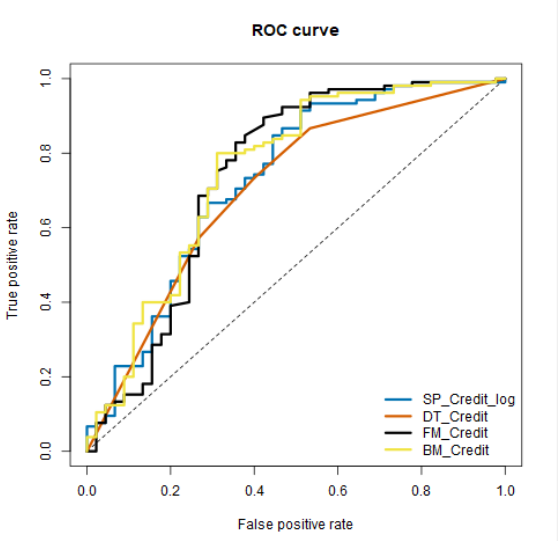
Relatório

Para predição de potenciais clientes como merecedores de crédito, foram testados os seguintes modelos preditivos de classificação binária: regressão logística, árvore de decisão, floresta e *boosted model*.

Foi escolhido o Modelo de Floresta, com precisão geral de 80%, e específica de “Creditworthy” de 96,19% e de “Non-Creditworthy” de 42,22%. A maior precisão de “Non-Creditworthy” foi alcançada pelo modelo de regressão logística – 48,89% -, porém, como a precisão de “Creditworthy” de 87,62% foi bem inferior à obtida no Modelo de Floresta, este teve resultado final melhor.



O gráfico ROC também revela o melhor desempenho do Modelo de Floresta. Nesse gráfico, pontos acima da diagonal (classificação completamente aleatória) representam bons resultados e, quanto mais afastados dessa linha (menos falsos negativos e falsos positivos), melhor.

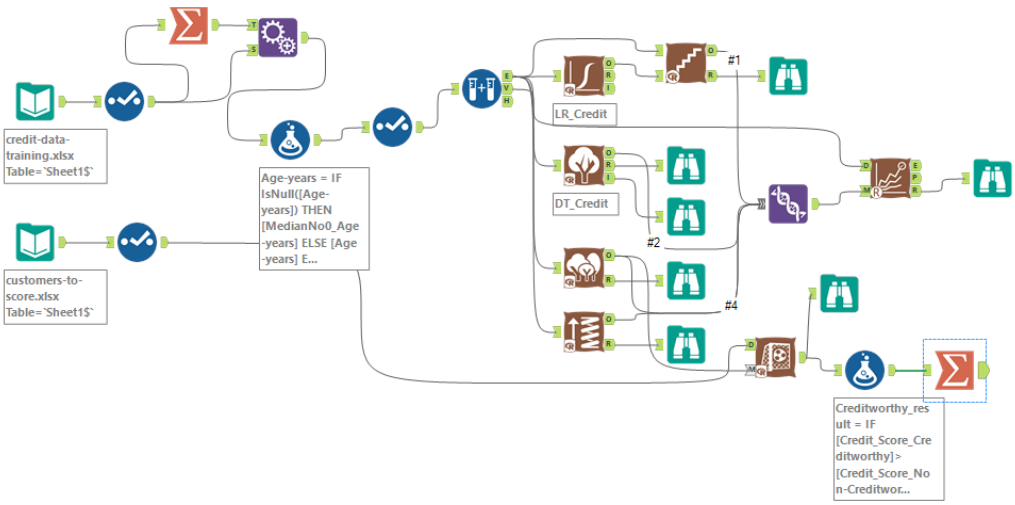


A análise das matrizes de confusão evidencia altas incidências de falsos positivos, o que revela considerável dificuldade na predição dessa categoria. A menor incidência foi a do Modelo de Regressão Logística (23), porém, como a ocorrência de falsos negativos é menor no caso do Modelo de Floresta (4 *versus* 13), este é o melhor modelo preditivo.

Ressalte-se: esse viés significa que clientes potencialmente inadimplentes podem ter seu empréstimo aprovado, o que, por sua vez, poderá resultar em eventuais prejuízos futuros.

Para a predição dos clientes confiáveis, foi considerada a seguinte fórmula: se Score_Creditworthy maior que Score_NonCreditworthy, a pessoa é "Creditworthy". Em consequência, a aplicação do modelo aos dados dos potenciais clientes resultou em um total de 408 aptos a receberem crédito do banco.

Fluxo no Alteryx:



Fontes:
https://es.wikipedia.org/wiki/Curva_ROC
http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/